Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models

Abstract

Recent text-to-image generative models have demonstrated an unparalleled ability to generate diverse and creative imagery guided by a target text prompt. While revolutionary, current state-of-the-art diffusion models may still fail in generating images that fully convey the semantics in the given text prompt. We analyze the publicly available Stable Diffusion model and assess the existence of catastrophic neglect, where the model fails to generate one or more of the subjects from the input prompt. Moreover, we find that in some cases the model also fails to correctly bind attributes (e.g., colors) to their corresponding subjects. To help mitigate these failure cases, we introduce the concept of Generative Semantic Nursing (GSN), where we seek to intervene in the generative process on the fly during inference time to improve the faithfulness of the generated images. Using an attention-based formulation of GSN, dubbed Attend-and-Excite, we guide the model to refine the cross-attention units to attend to all subject tokens in the text prompt and strengthen --- or excite --- their activations, encouraging the model to generate all subjects described in the text prompt. We compare our approach to alternative approaches and demonstrate that it conveys the desired concepts more faithfully across a range of text prompts. Code is available at our project page: https://attendandexcite.github.io/Attend-and-Excite/.